Data-Driven Hyperparameter Optimized Extreme Gradient Boosting Machine Learning Model for Solar Radiation Forecasting

Mantosh KUMAR[1], *Kumari NAMRATA*[1], *Nishant KUMAR*[2]

[1]Department of Electrical Engineering, National Institute of Technology, Adityapur, 831014 Jamshedpur, Jharkhand, India
[2]Department of Electrical Engineering, B. K. Birla Institute of Engineering & Technology, BKBIET Campus, CEERI Road, 333031 Pilani, Rajasthan, India

mantosh.nith@gmail.com, namrata.ee@nitjsr.ac.in, krnishant125@gmail.com

DOI: 10.15598/aeee.v20i4.4650

Article history: Received Jul 31, 2022; Revised Sep 07, 2022; Accepted Nov 23, 2022; Published Dec 31, 2022.
This is an open access article under the BY-CC license.

Abstract. *The uncertainty of the non-conventional sources especially solar energy caused due to spatio-temporal factors like temperature, pressure, relative humidity etc. is continuously disrupting the productivity and reliability of an integrated power system which motivates the researcher or energy industry for strategic forecasting solutions to enhance the proper scheduling and control of solar generation power plants. Several studies have been carried out; but still the objective of achieving accurate forecasting dependent on the spatio- temporal features is not achieved. To address this critical forecasting issue in this research article a hyper parametric tuning of the Extreme Gradient Boosting (XGB) machine learning model has been carried out using two met heuristic algorithms: Moth Flame Optimization (MFO) and Grey Wolf Optimization (GWO). The dataset comprises five years of metrological attributes collected from the National Renewable Energy Laboratory (NREL) for analysis. The validation of the proposed model has been done based on the five statistical errors: Max Error (ME), Mean Absolute Error (MAE), Coefficient of Determination (R^2), Mean Square Error (MSE) and Root Mean Square Error (RMSE). The regressive assessment of all three models has confirmed that the XGB-MFO model outperformed the others as showing the highest R^2 score of 0.9337, 0.9011, 0.8744 and lowest RMSE values of 76.29 $W{\cdot}m^{-2}$, 41.90$W{\cdot}m^{-2}$ and 95.94$W{\cdot}m^{-2}$ for Global Horizontal Irradiance (GHI), Diffuse Horizontal Irradiance (DHI) and Direct Normal Irradiance (DNI) respectively which ensures the proposed model implementation for the prediction and production of solar power.*

Keywords

Extreme Gradient Boosting, forecasting, Grey Wolf Optimization, Moth Flame Optimization, solar irradiance.

1. Introduction

1.1. Motivation

The consistent availability of energy supply across the nation is essential to a nation's economic prosperity [1]. With the rapid development of technology and urbanization [2], the need for a stable power supply is also increasing proportionally, pushing the power industry to complete shift on the Renewable Energy Sources (RES) in the long run-in order to fulfil the rising power consumption and reduce the greenhouse effect. As per recent International Energy Agency (IEA) analysis, the amount of energy produced via renewable sources surpassed 8,000 TWh in 2021, a record 500 TWh more than in 2020. At the same time hydropower decreased by 15 TWh, and wind and solar Photovoltaic (PV) output climbed by 270 TWh and 170 TWh, respectively.

The growth in worldwide CO_2 emissions in 2021 would have been 220 Mt greater without increased output from nuclear and renewable energy sources [13]. The total installed and pipelined solar capacity of India has been shown in Fig. 1 which depicts the progressive behaviour of the solar energy system [14]. Though the PV system is paving the path for clean energy, its intermittent nature makes its performance highly reliant on the weather and environment [15] and [16]. For the steady and secure integration of green energy sources into the present energy network accurate forecasting techniques have become essential [17] and [18]. Numerical Weather Prediction (NWP), statistical and Machine Learning (ML), and image-based methods are the three primary categories of solar forecasting techniques [19]. The NWP studies the forecasts of irradiance and weather while image-based approaches track and advection clouds using sky cameras, satellite photos, or shadow cameras to anticipate solar irradiance. Statistical and machine learning models 'train' themselves using past data and makes forecasts based on new input variable values. Statistical and ML methods can be used to a variety of time spans, but they are mostly used in hourly forecasting studies. Authors in [20] give an overview of trends in solar forecasting techniques.

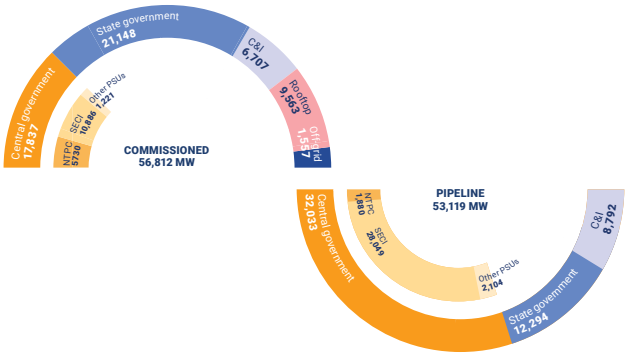

Fig. 1: Total installed and pipelined capacity by 31st December 2021, in (MW) [2].

Several researchers have proposed various ML models with default hyperparametric values which provide different prediction outcomes. In the present scenario, the hybridisation of metaheuristic algorithms with the ML model is being carried out to improve the accuracy for the various range of hyperparameters [21] and [22]. Although XGB has a considerably good performing model, the parametric search is essential for the development of the basic structure of any ML model which can be performed by incorporating the optimization methods. To address this hyperparametric search and develop an effective ML model for accurate solar irradiance forecasting, in this research article the hyperparametric tuning of the basic model of the XGB regressor has been performed by hybridization of the two optimization algorithms namely moth-flame optimization and grey wolf optimization method based on the 5 years dataset taken from NREL.

1.2. Status Quo of Solar Forecasting Using AI

Over the last few decades, many attempts have been made to forecast Solar Radiation using different sorts of empirical models, such as cloudiness-based models [23], sunshine-based models [24], and hybrid models that estimate global solar radiation by incorporating other meteorological factors. It has been predicted using ANN [25] and SVM [26] and the recent studies have been tabulated in Tab. 1.

1.3. Contributions to the Paper

The main objectives of our study are:

1. To create and analyze the XGB model for forecasting solar radiation utilizing web-based data, including.
2. To optimize the hyper parameters of the XGB model using MFO and GWO algorithms.
3. To compare all the three machine learning models accuracies and to find the best model among them for solar forecasting.

The remaining section of the research article has been structured as follows: Sec. 2. illustrates the dataset used and proposed methodology applied while the description of the algorithms incorporated has been briefly explained in Sec. 3. Section 4. describes the performance metrics used for the determination of the best model and Sec. 5. briefly explains the overall result analysis of all the ML models. Finally, the article has been concluded with the future scope in Sec. 6.

2. Methodology

2.1. Site Selection

As per the City Mayors Foundation, Jamshedpur, with coordinates (22°47'33 "N, 86°11'03 "E) is the 84th fastest-rising city globally. The Indian Meteorological Department Centre in Ranchi reports that Jamshedpur was the state's hottest location in 2022, with a scorching temperature of up to 43 °C [27]. Temperatures range from a minimum of 5 °C in winter to a maximum of about 43 °C in summer, and the average temperature of Jamshedpur is 25.7 °C [28].

Tab. 1: Literature survey of latest solar forecasted methods.

Authors	Objective	Solution
Jebli et al. [3]	ML models with the Pearson coefficient used to predict the real-time and shot time solar power.	Linear Regression (LR), Support Vector Regression model (SVR), Random Forest (RF), Multilayer Perceptron (MLP)
Kumari et al. [4]	An ensemble XGB-DNN method proposed for the estimation of the GHI on an hourly basis	Extreme Gradient Boosting (XGB), Deep Neural Network (DNN)
Trizoglou, et al. [5]	Authors have applied the XGB and LSTM in association with the SCADA system of wind turbines for forecasting of faults and reduce the operation and maintenance cost.	XGB, Long Short-Term Memory (LSTM)
Lee et al. [6]	An ensemble technique for forecasting the solar radiation for a short duration used which shows more reliable outputs as compared to individual ML models.	Ensemble Method Bagged-Trees, Boosted-Trees, RF, Support Vector Machines (SVM), Gaussian Process Regression (GPR)
Massaoudi et al. [7]	A Stacking method used to combine the three ML models (XGB-LGBM-MLP) to forecast the grid load for short duration.	Stacking (XGB-LGBM-MLP)
Mokbal et al. [8]	Extreme Gradient Boosting Cross-Site Scripting (XGBXSS) method used for detecting the Cross-Site Scripting attacks where XGB has been applied with the feature selection and a recursive optimization.	XGB, Grid Search
Fan et al. [9]	ML models were used to predict the transpiration of daily maize and it was concluded that the DNN model is more efficient for daily maize T estimate.	XGB, Artificial Neural Networks (ANN), DNN, SVM
Nguyen et al. [10]	XGB applied to forecast the punching shear resistance of R/C interior slabs. The designed XGB model's prediction accuracy for punching shear strength was investigated and compared to other machine learning models and empirical models.	XGB, ANN, RF
Chia et al. [11]	XGB with met heuristic models i.e. MFO, Whale Optimization Algorithm (WOA) and Particle Swarm Optimization (PSO) have been used for evapotranspiration estimation	XGB with PSO, MFO, WOA
Rui Liu et al. [12]	GWO has been incorporated with ML models for groundwater potential prediction.	GWO with RF and SVM

2.2. Data Pre-processing

The meteorological data is in its raw state and must be pre-processed before it can be used. In the data pre-processing, there is a combination of the following four processes. Figure 2 depicts the workflow for our experiment.

1. Data Cleaning: It involves checking for repeated, duplicate, and Not Applicable (NA) entries in the data.

2. Data Normalization: Here, all data variables are normalized to a common interval, which is often between 0 and 1. This phase compares the values of numerous variables.

3. Feature Extraction: Here, only important features are picked, as including unnecessary features increases data size and slow down a forecasting algorithm's learning speed and accuracy. It is done through Exploratory Data Analysis (EDA) process.

4. Data Splitting: The pre-processed dataset is divided into test and training sets and sent to the forecasting phase.

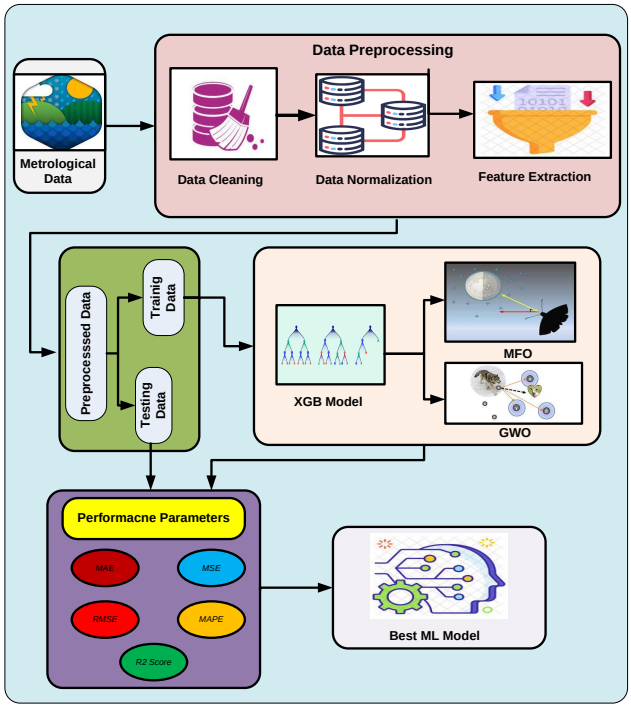

Fig. 2: Methodology of the proposed work.

3. Forecasting Algorithm Used

3.1. Extreme Gradient Boosting (XGB) ML Model

Chen and Guestrin developed the XGB algorithm as a revolutionary implementation approach for Gradient Boosting Machines, namely Regression Trees and K Classification [29]. XGB is designed to avoid over fitting and optimizing computation resources at the same time. During the training phase of XGB, calculations are also performed synchronously and automatically for all the functions. The model's final prediction is calculated as the sum of each model's predictions. The pseudo-code description for XGB is given by Algorithm 3 and the schematic diagram for the XGB algorithm is shown in Fig. 3.

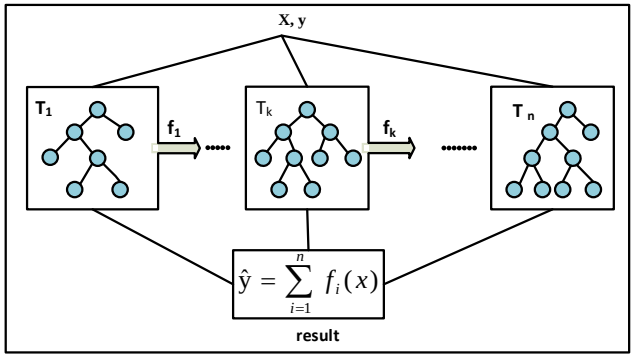

Fig. 3: XGB model.

3.2. Moth Flame Optimization (MFO)

In 2015, author in [30] proposed the MFO algorithm, which was motivated by the mirroring behaviour of moths. These moths employ a peculiar kind of nocturnal triangulation known as transverse orientation, which allows moths to hover in a straight line by remembering the stationary perspective parallel to the moon. Moths float in spiral patterns in the latency of an unreal source of light that is close to the moon by focusing upon the source of light.

$$AM = \begin{pmatrix} AM_1 \\ AM_2 \\ \vdots \\ AM_a \end{pmatrix}. \tag{5}$$

Moths and flames are two significant components of the MFO structure. The moths that hover in a deeply engaged, d-dimensional plane act as search mediators. In the **M** matrix, the dwelling is reserved. The fitness value relevant to each month is subsequently stored in array AM. The size of a moth

Algorithm 1 Implementation of XGB.

Input: Dataset D, X (*Features*) and y (*Target*) loaded with training labelled data, parameters (estimators, learning rate, maximum depth etc.).
Output: System accuracy in terms of performance metrics.
Initialize a base model with:

$$f_0(x) = \arg\min_{\gamma} \sum_{i=1}^{m} L(y_i, \gamma) + \Omega. \tag{1}$$

while (stopping criterion) **do**
  for $t = 1, t++$ **do**
    for $i = 1$ to m **do**
      Compute residual, r_{it}:

$$r_{it} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f-f_{i-1}}. \tag{2}$$

    end for
    for $j = 1$ to J_t **do**
      Fit the weak tree to r_{it}.
      Compute revised loss function:

$$\gamma_{jt} = \arg\min_{\gamma} \sum_{x_i} L(y_i, f_{t-1}(x_i) + \gamma). \tag{3}$$

      Update model function:

$$f_t(x) = f_{t-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm}). \tag{4}$$

    end for
  end for
end while
Model fitting with training data.
Model validation with testing data.

and a flame are the same. Moth and flame both function as parts of the algorithmic solution. Flame denotes the moth's ideal position, whereas the moth denotes the hunting agent. Moths revolves around the flames that serve as flag throughout the search process. As a result, both positions are being updated, decreasing the likelihood that one would be lost. According to Eq. (6), the moth's location is updated.

$$M_j = SF(M_j, F_j), \tag{6}$$

where M_j indicates the j^{th} moths, whereas F_j represents the j^{th} flames and SF is for spiral function which is expressed in Eq. 7.

$$SF(M_j, F_j) = D_j{}^* e^{bt*} \cos(2\pi t) + F_j, \tag{7}$$

where, b is spiral constant, t is the arbitrary value $(-1,1)$ and D_j is j^{th} moth and j^{th} flame Euclidean

distance. D_j is represented as Eq. (8).

$$D_j = |F_j - M_j| . \tag{8}$$

In the initial stage, flames and moths remain to be the exact number, which may reduce the potential of sophisticated solutions to be diverse due to moths' conscious choice of n distinct locations in the quest of room for updating. Eq. (9) is used to update the flames.

$$F_{no} = \text{round}\left(F - j^* \frac{F-1}{ir_{\max}} O\right). \tag{9}$$

3.3. Grey Wolf Optimization (GWO)

GWO, modelled on the natural hunting tactics of grey wolves is a meta-heuristic optimization technique proposed in 2014 by authors in [31]. Every wolf in GWO symbolizes a search agent (potential solution). GWO classifies the wolves into four categories alpha (α), beta (β), gamma (δ) and omega (ω) by replicating the grey wolf population's hierarchy. The wolves in the first three grades correspond to the current three best solutions. The current three best solutions are represented by the wolves in the first three categories (α, β, δ). The (ω) wolves follow the pack's strongest wolves.

1) Encircling

Grey wolves surround their prey as part of the hunting process. So, the initial phase of the mathematical modelling of the GWO is to surround the target, which may be expressed by the following formulas [16] and [17].

$$\vec{D}_{GW} = \left|\vec{C} \cdot \vec{X}^p_{GW}(t) - \vec{X}_{GW}(t)\right|, \tag{10}$$

$$\vec{X}_{GW}(t+1) = \left|\vec{X}^p_{GW}(t) - \vec{A} \cdot \vec{D}\right|, \tag{11}$$

where, $\vec{A}$ and $\vec{C}$ are noted as the coefficient vectors and t is symbolised as the current iterations. $\vec{X}_{GW}$ signifies grey wolf position vector and Prey's position vector is indicated by $\vec{X}^p_{GW}$ whereas, the $\vec{D}_{GW}$ is the vector which depends on $\vec{X}^p_{GW}$. Computation for the coefficient vectors $\vec{A}$ and $\vec{C}$ are as follows:

$$\vec{A} = 2\vec{a} \cdot \vec{r}_1 - \vec{a}, \tag{12}$$

$$\vec{C} = 2 \cdot \vec{r}_2, \tag{13}$$

$$\vec{a} = 2 - 2\left(\frac{itr}{itr_{\max}}\right), \tag{14}$$

where, $\vec{r}_1$ and $\vec{r}_2$ are random variables in the interval [0, 1] and values of $\vec{a}$ are linearly decreasing from 2 to 0 throughout the span of iterations. Concisely, $\vec{r}_1$ and $\vec{r}_2$ vectors enable wolves to extend to any location. Accordingly, Eq. (13) and Eq. (14) indicates that the grey wolf may update their position inside the search space (space circling prey) at any random point. The same approach could be employed in a search space with dimension n, where the grey wolves will circle the best outcome thus far in hyper-cubes or hyper-spheres.

2) Hunting

The α usually leads the hunt while the β and δ may occasionally engage in hunting. We postulate that the alpha (best solution), beta, and delta have superior information about the probable location of prey to mathematically imitate the hunting behaviour of grey wolves. Therefore, we reserve the first three best responses. Thus, to compel the other searching agent, along with omegas and to upgrade their positions in accordance with the status of the top search agents. The below mentioned Eq. (15), Eq. (16), Eq. (17) and Eq. (18) are followed for above stated context:

$$\vec{X}_{GW}(t+1) = \frac{\vec{X}^1_{GW} + \vec{X}^2_{GW} + \vec{X}^3_{GW}}{3}, \tag{15}$$

$$\vec{X}^1_{GW} = \vec{X}^\alpha_{GW} - \vec{A}_1 \cdot \left(\left|\vec{C}_1 \cdot \vec{X}^\alpha_{GW} - \vec{X}_{GW}\right|\right), \tag{16}$$

$$\vec{X}^2_{GW} = \vec{X}^\beta_{GW} - \vec{A}_2 \cdot \left(\left|\vec{C}_2 \cdot \vec{X}^\beta_{GW} - \vec{X}_{GW}\right|\right), \tag{17}$$

$$\vec{X}^3_{GW} = \vec{X}^\delta_{GW} - \vec{A}_3 \cdot \left(\left|\vec{C}_3 \cdot \vec{X}^\delta_{GW} - \vec{X}_{GW}\right|\right). \tag{18}$$

3) Searching and Attacking Prey

Grey wolves primarily use the (α), (β), and (δ) positions to guide their search. They disperse from one another to look for prey and then reassemble to attack it. We use $\vec{A}$ with random values higher than 1 or less than -1 to force the search agent to diverge from the prey to mathematically simulate divergence. This encourages exploration and enables a wide search for the GWO algorithm. As already mentioned, after the prey stops moving, the grey wolves attack it to end the hunt. We lower the value of $\vec{a}$ to mathematically simulate approaching the prey. Keeping in mind the reduction occurring in $\vec{A}$ may also decrease by $\vec{a}$. In other respects, $\vec{a}$ decreases from 2 to 0 throughout the duration of iterations, and $\vec{A}$ is a random number in the range [2a, 2a]. $\vec{A}$ search agent's future position may be anywhere between its present position and the prey's position when random numbers of $\vec{A}$ are in the range [1, 1].

4. Performance Parameters

To quantify the performance and their variation from the real value for the ML models, we provide

Algorithm 2 Pseudo code of GWO.

Input: Wolf Population (N), $\vec{A}$, $\vec{a}$, and $\vec{C}$.
Output: Optimal solution (R_2).
Fitness Calculation of search agent (i.e Grey wolves).
$\vec{X}^{\alpha}_{GW}$ best optimal solution (search agent).
$\vec{X}^{\beta}_{GW}$ second best optimal solution (search agent).
$\vec{X}^{\delta}_{GW}$ third best optimal solution (search agent).
while $(itr < itr_{\max})$ **do**
  for $i = 1, 2, 3, \ldots N$ **do**
    Update current position using Eq. (18).
  end for
  Update $\vec{A}$, $\vec{a}$, and $\vec{C}$.
  Fitness Calculation of search agent.
  Update $\vec{X}^{\alpha}_{GW}$, $\vec{X}^{\beta}_{GW}$ and $\vec{X}^{\delta}_{GW}$.
  $itr = itr + 1$.
end while
return $\vec{X}^{\alpha}_{GW}$.

several common statistical metrics. The difference between the estimated (or anticipated) and actual output parameter is known as the deviation sometimes referred to as the errors or residue. For example, the error for GHI can be expressed as:

$$\delta = GHI_{obs} - GHI_{\text{pred}} \,. \tag{19}$$

These can be used to assess the degree of divergence and correlation between the predicted and actual data. Figure 4 depicts the expressions for forecasting the effectiveness of ML models in our research.

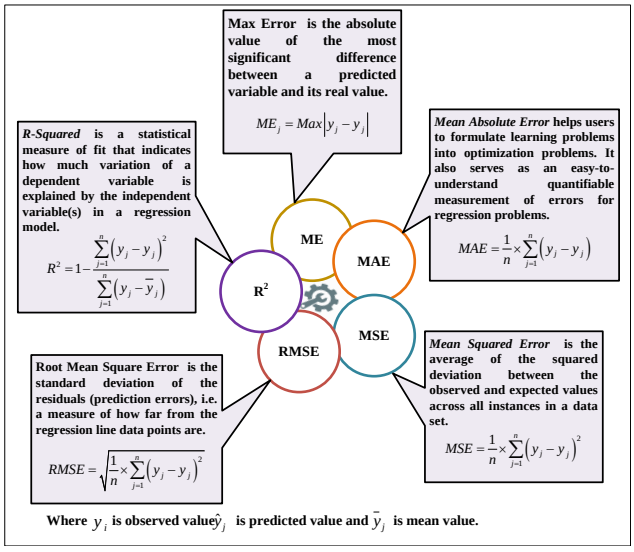

Fig. 4: Performance Evaluation parameter used in our work.

5. Results and Discussion

The objective of the study is to develop an optimized system for forecasting solar irradiance using the XGB model for the selected region, as well as two optimization techniques have been incorporated to optimize the parameters of the XGB to enhance the performance of prediction. Numerous research papers have been published about the study of this kind of model. Due to the complexity of the time series and the accumulation of forecasting mistakes, it is still difficult to determine how to best optimize the XGB model, using the met heuristic optimization techniques for the prediction of solar irradiation.

Hence, two hybrid model XGB-MFO and XGB-GWO have been analysed for the forecasting purpose. Initially the 70 % of dataset i.e., training data has been used for the training the two hybrid models and the functions built into the system are evaluated by comparing the forecasted results to the real outcomes based on statistical errors. This validates the recommended methodological approach used. On the validation dataset, which includes 30 % of the dataset, five evaluation metrics MAE, MSE, RMSE, ME and R^2 score is utilised to determine the best hyper parametric optimized model. The model with the lowest error and highest accuracy is finally noted as the best predictive model.

It's crucial to keep in mind that the population size of each hybrid model varies and changing this parameter's value will have an immediate impact on the model's running duration and ability to identify the overall best solution. A large population will greatly lengthen the running time, which will make it difficult to apply the models to engineering problems, while a small population would result in unstable fitness values. Five population sizes - 50, 100, 150, 200 and 500 were used in this study to construct the two hybrid models.

The whole system has been designed using Python language where the system has been trained for 3 years and validated for the next 1 year i.e., 35,078 entries have been used for model training and 4,922 entries applied for validating the model to obtain the optimum result.

5.1. Performance of XGB ML Model

Tab. 2: Selected values for XGB Hyper-parameters.

Sl. No.	Description	Value
1	Maximum no. of trees	100
2	Maximum depth	5
3	Learning rate	0.001

Tab. 3: Performance evaluation parameters outcomes for XGB without optimization.

	MAE ($W \cdot m^{-2}$)	MSE ($W \cdot m^{-2}$)	RMSE ($W \cdot m^{-2}$)	ME ($W \cdot m^{-2}$)	R^2 score	
					Train	Test
GHI	38.144	6449.03	80.30	791.74	0.9720	0.9278
DHI	25.265	2317.48	48.14	363.31	0.9611	0.8751
DNI	50.471	9318.99	96.53	692.92	0.9189	0.8527

Generally subsampling happens once for every tree in XGB. Increasing the depth of the tree makes the model more complicated and prone to over fitting.

To prevent over fitting, step size shrinking is employed in the weight update with the help of learning rate. We may immediately obtain the new weights of the features after each boosting step, and the learning rate, here, lowers the weights of the feature to make the boosting method more conservative. The parameters selected for the analysis the XGB model without optimization has been shown in Tab. 2 where hyper parameters has been fixed.

The evaluation parameters obtained using the selected hyper parametric values have been tabulated in Tab. 3 which shows the higher error values for DNI than GHI and DHI while the R^2 score of GHI is approximately 8 % higher as compared to the DHI and DNI parameter. The efficiency of the model can be improved using the proper parameters of the XGB model.

To optimize the model, the range of various internal parameters of the XGB has been taken as shown in Tab. 4 which will be optimized using the two optimization methods i.e., MFO and GWO. The iteration for both optimization methods has been fixed to 100 to analyse the effect of increasing the population size. The evaluation parameters for determining the effect of the MFO and GWO optimization on the XGB model has been shown in Tab. 5 and Tab. 6, where all the five parameters for the target variable GHI, DHI and DNI have been calculated for the various population size to check the best population of the nature-based algorithm with respect to the fixed iteration count. The best values for each parameter have been highlighted in bold which shows that the model has been optimized as the statistical errors have been reduced and the accuracy of the model has been considerably improved. The graphical representation of the accuracy of the XGB-MFO model for various population sizes has been shown in Fig. 5.

Tab. 4: Selected values for XGB hyper-parameters.

Sl. No.	Description	Upper bound	Lower bound
1	Maximum no. of trees	100	1000
2	Maximum depth	5	50
3	Learning rate	0.001	0.1

Additional testing of the novel approach will be necessary for data derived from weather forecasts. However, the model must employ predicted weather information to be useful. The model would be especially helpful for applications on a wider scale (i.e., county, or regional scale). This is possible due to the ability to smooth out the quick change in local meteorological conditions that causes the intra-hourly fluctuation in solar irradiance. Larger-scale PV output tests in conjunction with anticipated weather data encourage the proper use of the proposed model.

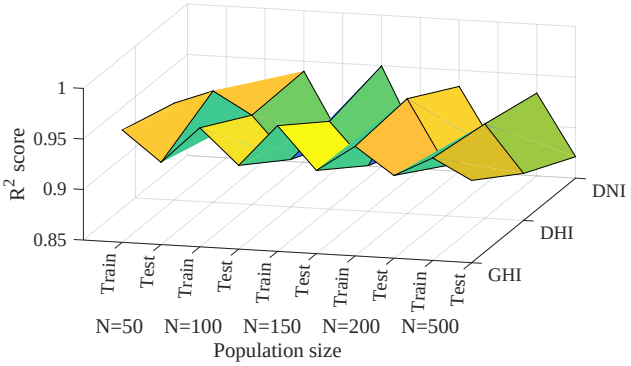

Fig. 5: R^2 score analysis using XGB-MFO (100 iterations).

5.2. Comparative Analysis

In this section, all three models i.e., XGB, XGB-MFO and XGB-GWO are compared with the best parameters to identify the well-suited ML model for forecasting solar irradiance. The overall best comparative performance analysis has been represented in Tab. 7, which shows that the XGB-GWO has an accuracy of 0.63 %, 2.88 % 2.48 % and XGB-MFO has 0.53 %, 2.73 % and 2.40 % accuracy more than that of the unoptimized XGB model for GHI, DHI and DNI respectively. The RMSE values have also been reduced by 4.98 %, 12.94 %, 0.61 % using XGB-GWO and 4.35 %, 12.31 % and 0.30 % using XGB-MFO predicted model for the three target parameters which clearly signifies the contribution the meta-heuristic algorithms with the ML models. The corresponding population sizes were 100, 200 and 150 which shows the importance of the proper selection of population sizes. The best outcomes in the table have been highlighted. Overall, we can state that XGB-GWO outperforms the other two models for the given datasets and location. The comparative analysis has also been represented in Fig. 6.

Tab. 5: Performance evaluation parameters outcomes for XGB-MFO.

	Iterations	Population size	MAE ($W \cdot m^{-2}$)	MSE ($W \cdot m^{-2}$)	RMSE ($W \cdot m^{-2}$)	ME ($W \cdot m^{-2}$)	R^2 score	
							Train	Test
GHI	100	50	33.8953	5947.45	77.0114	726.15	0.9611	0.9314
		100	**33.4033**	**5899.41**	**76.8076**	**738.05**	**0.9678**	**0.9328**
		150	32.8728	5939.38	77.0674	748.90	0.9744	0.9323
		200	33.2459	5983.11	77.3505	743.24	0.9584	0.9318
		500	33.2469	6023.76	77.6128	750.95	0.9514	0.9314
DHI	100	50	20.1519	1841.72	42.9153	361.75	0.9459	0.8963
		100	20.0928	1834.19	42.8274	364.53	0.9384	0.8968
		150	20.8935	1859.86	43.1260	361.86	0.9368	0.8953
		200	**20.7235**	**1781.66**	**42.2097**	**349.77**	**0.9637**	**0.8997**
		500	20.1217	1838.93	42.8827	365.12	0.9432	0.8965
DNI	100	50	48.7757	9258.24	96.2197	683.43	0.9602	0.8736
		100	49.0448	9276.45	96.3143	694.41	0.9401	0.8735
		150	**48.9292**	**9262.21**	**96.2404**	**693.43**	**0.9498**	**0.8737**
		200	49.3427	9377.84	96.8392	699.48	0.9340	0.8721
		500	49.6549	9446.23	97.1917	697.78	0.9319	0.8712

Tab. 6: Performance evaluation parameters outcomes for XGB-GWO.

	Iterations	Population size	MAE ($W \cdot m^{-2}$)	MSE ($W \cdot m^{-2}$)	RMSE ($W \cdot m^{-2}$)	ME ($W \cdot m^{-2}$)	R^2 score	
							Train	Test
GHI	100	50	33.4417	5823.47	76.3117	705.48	0.9822	0.9336
		100	**33.0237**	**5821.60**	**76.2994**	**685.70**	**0.9884**	**0.9337**
		150	33.70866	5906.59	76.8543	739.83	0.96533	0.9327
		200	32.7415	5944.24	77.0988	747.75	0.9613	0.9323
		500	32.97887	5978.46	77.3205	747.09	0.9604	0.9319
DHI	100	50	19.8547	1805.28	42.4886	361.32	0.9651	0.8984
		100	19.9677	1829.09	42.7679	363.31	0.9481	0.8979
		150	20.74984	1843.34	42.9342	356.59	0.9441	0.8962
		200	**20.46198**	**1756.35**	**41.9088**	**352.06**	**0.9788**	**0.9011**
		500	20.1074	1785.00	42.2493	357.94	0.9613	0.8995
DNI	100	50	48.5425	9058.75	95.1774	673.65	0.9674	0.8739
		100	48.9012	9241.05	96.1304	687.33	0.9479	0.8740
		150	**48.8938**	**9205.37**	**95.9446**	**671.52**	**0.9612**	**0.8744**
		200	48.9643	9244.79	96.1498	690.55	0.9434	0.8739
		500	49.0030	9392.69	96.9159	692.02	0.9308	0.8719

Tab. 7: XGB models comparative analysis.

	ML models	MAE ($W \cdot m^{-2}$)	MSE ($W \cdot m^{-2}$)	RMSE ($W \cdot m^{-2}$)	ME ($W \cdot m^{-2}$)	R^2 score	
						Train	Test
GHI	XGB	38.1442	6449.03	80.3058	791.74	0.9720	0.9278
	XGB-MFO	33.4033	5899.41	76.8076	738.05	0.9678	0.9328
	XGB-GWO	**33.0237**	**5821.60**	**76.2994**	**685.70**	**0.9884**	**0.9337**
DHI	XGB	25.2654	2317.48	48.1402	363.31	0.9611	0.8751
	XGB-MFO	20.7235	1781.66	42.2097	349.77	0.9637	0.8997
	XGB-GWO	**20.46198**	**1756.35**	**41.9088**	**352.06**	**0.9788**	**0.9011**
DNI	XGB	50.4714	9318.99	96.5349	692.92	0.9189	0.8527
	XGB-MFO	48.9292	9262.21	96.2404	693.43	0.9498	0.8737
	XGB-GWO	**48.8938**	**9205.37**	**95.9446**	**671.52**	**0.9612**	**0.8744**

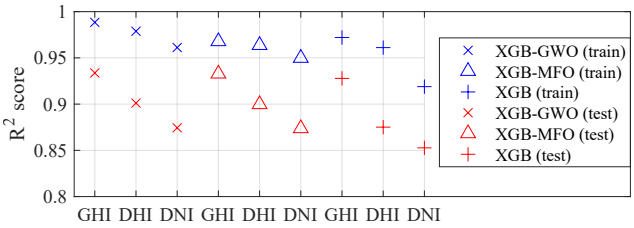

Fig. 6: Comparative analysis of all models based on R^2 score.

6. Conclusion

The prediction of accurate solar irradiance is very useful for the forecasting of solar energy. The main goal of these techniques is to modify the XGB's combination of hyper parameters using most prominent optimization algorithms, such as GWO, and MFO, to increase the prediction accuracy which can be useful for engineering practise. As a result, in this study, the XGB-MFO and XGB-GWO hybrid models have been created. Five statistical parameters were chosen to assess the consistency between the actual value and the fore-

casted value to examine the performance of each hybrid model. An unoptimized XGB model was also developed, verified, and trained using the NREL historical datasets to evaluate the performance of the two optimization techniques. The experimental findings show that both in the training stage and the test stage, the two XGB-based hybrid models performed much better than the unoptimized XGB model.

The two hybrid models' prediction accuracy exceeded 0.9 during testing, particularly the XGB-GWO model (for GHI, R^2 score: 0.9337; MSE: 5821.60; RMSE: 76.2994; ME: 685.70; MAE: 33.0237), whose prediction accuracy reached 0.93 which ensures the proposed model applicability for the further forecasting purpose.

Author Contributions

M.K. has collected and analysed the data along with computation and mathematical modelling for the methodology adopted. K.N. supervised the project and formatted the manuscript. N.K. has worked in optimization technique formulation and assisted in editing and formulating the manuscript.

References

[1] GUO, Z., K. ZHOU, C. ZHANG, X. LU, W. CHEN and S. YANG. Residential electricity consumption behavior: Influencing factors, related theories and intervention strategies. *Renewable and Sustainable Energy Reviews*. 2018, vol. 81, iss. 1, pp. 399–412. ISSN 1364-0321. DOI: 10.1016/j.rser.2017.07.046.

[2] ZHANG, Y., C.-Q. HE, B.-J. TANG and Y.-M. WEI. China's energy consumption in the building sector: A life cycle approach. *Energy and Buildings*. 2015, vol. 94, iss. 1, pp. 240–251. ISSN 0378-7788. DOI: 10.1016/j.enbuild.2015.03.011.

[3] JEBLI, I., F.-Z. BELOUADHA, M. I. KABBAJ and A. TILIOUA. Prediction of solar energy guided by pearson correlation using machine learning. *Energy*. 2021, vol. 224, iss. 1, pp. 1–20. ISSN 0360-5442. DOI: 10.1016/j.energy.2021.120109.

[4] KUMARI, P. and D. TOSHNIWAL. Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance. *Journal of Cleaner Production*. 2021, vol. 279, iss. 1, pp. 1–14. ISSN 0959-6526. DOI: 10.1016/j.jclepro.2020.123285.

[5] TRIZOGLOU, P., X. LIU and Z. LIN. Fault detection by an ensemble framework of Extreme Gradient Boosting (XGBoost) in the operation of offshore wind turbines. *Renewable Energy*. 2021, vol. 179, iss. 1, pp. 945–962. ISSN 0960-1481. DOI: 10.1016/j.renene.2021.07.085.

[6] LEE, J., W. WANG, F. HARROU and Y. SUN. Reliable solar irradiance prediction using ensemble learning-based models: A comparative study. *Energy Conversion and Management*. 2020, vol. 208, iss. 1, pp. 1–13. ISSN 0196-8904. DOI: 10.1016/j.enconman.2020.112582.

[7] MASSAOUDI, M., S. S. REFAAT, I. CHIHI, M. TRABELSI, F. S. OUESLATI and H. ABU-RUB. A novel stacked generalization ensemble-based hybrid LGBM-XGB-MLP model for Short-Term Load Forecasting. *Energy*. 2021, vol. 214, iss. 1, pp. 1–14. ISSN 0360-5442. DOI: 10.1016/j.energy.2020.118874.

[8] MOKBAL, F. M. M., W. DAN, W. XIAOXI, Z. WENBIN and F. LIHUA. XGBXSS: An Extreme Gradient Boosting Detection Framework for Cross-Site Scripting Attacks Based on Hybrid Feature Selection Approach and Parameters Optimization. *Journal of Information Security and Applications*. 2021, vol. 58, iss. 1, pp. 1–20. ISSN 2214-2126. DOI: 10.1016/j.jisa.2021.102813.

[9] FAN, J., J. ZHENG, L. WU and F. ZHANG. Estimation of daily maize transpiration using support vector machines, extreme gradient boosting, artificial and deep neural networks models. *Agricultural Water Management*. 2021, vol. 245, iss. 1, pp. 1–12. ISSN 0378-3774. DOI: 10.1016/j.agwat.2020.106547.

[10] NGUYEN, H. D., G. T. TRUONG and M. SHIN. Development of extreme gradient boosting model for prediction of punching shear resistance of r/c interior slabs. *Engineering Structures*. 2021, vol. 235, iss. 1, pp. 1–14. ISSN 0141-0296. DOI: 10.1016/j.engstruct.2021.112067.

[11] CHIA, M. Y., Y. F. HUANG and C. H. KOO. Swarm-based optimization as stochastic training strategy for estimation of reference evapotranspiration using extreme learning machine. *Agricultural Water Management*. 2021, vol. 243, iss. 1, pp. 1–15. ISSN 0378-3774. DOI: 10.1016/j.agwat.2020.106447.

[12] LIU, R., G. LI, L. WEI, Y. XU, X. GOU, S. LUO and X. YANG. Spatial prediction of groundwater potentiality using machine learning methods with Grey Wolf and Sparrow Search Algorithms. *Journal of Hydrology*. 2022,

vol. 610, iss. 1, pp. 1–20. ISSN 0022-1694. DOI: 10.1016/j.jhydrol.2022.127977.

[13] Global Energy Review: CO2 Emissions in 2021 – Analysis. In: *IEA* [online]. 2022. Available at: https://www.iea.org/reports/global-energy-review-co2-emissions-in-2021-2.

[14] India Solar Compass Q1 2022. In: *Bridge To India Private* [online]. 2022. Available at: https://bridgetoindia.com/report/india-solar-compass-q1-2022/.

[15] VAN DER MEER, D. W., M. SHEPERO, A. SVENSSON, J. WIDEN and J. MUNKHAMMAR. Probabilistic forecasting of electricity consumption, photovoltaic power generation and net demand of an individual building using Gaussian Processes. *Applied Energy*. 2018, vol. 213, iss. 1, pp. 195–207. ISSN 0306-2619. DOI: 10.1016/j.apenergy.2017.12.104.

[16] KUMAR, N., K. NAMRATA and A. SAMADHIYA. Bi-level decision making in techno-economic planning and probabilistic analysis of community based sector-coupled energy system. *Applied Intelligence*. 2022, pp. 1–25. ISSN 1573-7497. DOI: 10.1007/s10489-022-03794-9.

[17] Future of Photovoltaic. In: *International Renewable Energy Agency* [online]. 2019. Available at: https://www.irena.org/publications/2019/Nov/Future-of-Solar-Photovoltaic.

[18] SAMADHIYA, A., K. NAMRATA and N. KUMAR. An Experimental Performance Evaluation and Management of a Dual Energy Storage System in a Solar Based Hybrid Microgrid. *Arabian Journal for Science and Engineering*. 2022, pp. 1–24. ISSN 2191-4281. DOI: 10.1007/s13369-022-07023-w.

[19] YAGLI, G. M., D. YANG and D. SRINIVASAN. Automatic hourly solar forecasting using machine learning models. *Renewable and Sustainable Energy Reviews*. 2019, vol. 105, iss. 1, pp. 487–498. ISSN 1364-0321. DOI: 10.1016/j.rser.2019.02.006.

[20] YANG, D., J. KLEISSL, C. A. GUEYMARD, H. T. C. PEDRO and C. F. M. COIMBRA. History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining. *Solar Energy*. 2018, vol. 168, iss. 1, pp. 60–101. ISSN 0038-092X. DOI: 10.1016/j.solener.2017.11.023.

[21] KURNIABUDI, D. STIAWAN, DARMAWIJOYO, M. Y. B. IDRIS, S. DEFIT, Y. S. TRIANA and R. BUDIARTO. Improvement of attack detection performance on the internet of things with PSO-search and random forest. *Journal of Computational Science*. 2022, vol. 64, iss. 1, pp. 1–13. ISSN 1877-7503. DOI: 10.1016/j.jocs.2022.101833.

[22] KUMAR, N., K. NAMRATA and A. SAMADHIYA. Deterministic Robust Planning and Probabilistic Techno-Economic Assessment of a Sector Coupled Community Energy System. *Advanced Theory and Simulations*. 2022, vol. 5, iss. 5, pp. 1–24. ISSN 2513-0390. DOI: 10.1002/adts.202100639.

[23] BADESCU, V. A new kind of cloudy sky model to compute instantaneous values of diffuse and global solar irradiance. *Theoretical and Applied Climatology*. 2002, vol. 72, iss. 1–2, pp. 127–136. ISSN 1434-4483. DOI: 10.1007/s007040200017.

[24] YADAV, A. K. and S. S. CHANDEL. Solar radiation prediction using Artificial Neural Network techniques: A review. *Renewable and Sustainable Energy Reviews*. 2014, vol. 33, iss. 1, pp. 772–781. ISSN 1364-0321. DOI: 10.1016/j.rser.2013.08.055.

[25] CHEN, J.-L., G.-S. LI, B.-B. XIAO, Z.-F. WEN, M.-Q. LV, C.-D. CHEN, Y. JIANG, X.-X. WANG and S.-J. WU. Assessing the transferability of support vector machine model for estimation of global solar radiation from air temperature. *Energy Conversion and Management*. 2015, vol. 89, iss. 1, pp. 318–329. ISSN 0196-8904. DOI: 10.1016/j.enconman.2014.10.004.

[26] TAO, H., A. A. EWEES, A. O. AL-SULTTANI, U. BEYAZTAS, M. M. HAMEED, S. Q. SALIH, A. M. ARMANUOS, N. AL-ANSARI, C. VOYANT, S. SHAHID and Z. M. YASEEN. Global solar radiation prediction over North Dakota using air temperature: Development of novel hybrid intelligence model. *Energy Reports*. 2021, vol. 7, iss. 1, pp. 136–157. ISSN 2352-4847. DOI: 10.1016/j.egyr.2020.11.033.

[27] Comprehensive Clean Air Action Plan. In: *Central Pollution Control Board* [online]. 2021. Available at: https://cpcb.nic.in/Actionplan/Jamshedpur.pdf.

[28] Post-Monsoon Report-2022 (JHARKHAND). In: *Mausam* [online]. 2022. Available at: https://mausam.imd.gov.in/ranchi/mcdata/SWR_Jharkhand.pdf.

[29] YAN, S., L. WU, J. FAN, F. ZHANG, Y. ZOU and Y. WU. A novel hybrid WOA-XGB model for estimating daily reference evapotranspiration using local and external meteorological data: Applications in arid and humid regions of China. *Agricultural Water Management*. 2021,

vol. 244, iss. 1, pp. 1–22. ISSN 0378-3774. DOI: 10.1016/j.agwat.2020.106594.

[30] MIRJALILI, S. Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm. *Knowledge-Based Systems*. 2015, vol. 89, iss. 1, pp. 228–249. ISSN 0950-7051. DOI: 10.1016/j.knosys.2015.07.006.

[31] MIRJALILI, S., S. M. MIRJALILI and A. LEWIS. Grey Wolf Optimizer. *Advances in Engineering Software*. 2014, vol. 69, iss. 1, pp. 46–61. ISSN 0965-9978. DOI: 10.1016/j.advengsoft.2013.12.007.

About Authors

Mantosh KUMAR received the B.Tech. degree from Biju Patnaik University of Technology (BPUT), Rourkela, India, and M.Tech. degree from National Institute of Technology (NIT), Hamirpur, India. He is currently working as research scholar in Electrical Engineering NIT, Jamshedpur, India. His current research interests include Renewable Energy and Machine Learning.

Kumari NAMRATA (corresponding author) (Member, IEEE) is an associate professor in the Department of Electrical Engineering, NIT, Jamshedpur, Jharkhand. Her research interests include Solar Power Generation and Conversion, Solar radiation estimation and forecasting. She has published more than 60 scientific papers in reputed journals, books and international conferences related to the field on solar radiation forecasting, Solar based Distributed generation, modelling, and control. She is a reviewer of International Journal of Emerging Electric Power Systems, International Transactions on Electrical Energy Systems.

Nishant KUMAR is an assistant rofessor in the department of Electrical Engineering, B. K. Birla Institute of Engineering and Technology, Pilani, Rajasthan, India. He has submitted the Ph.D. thesis at NIT Jamshedpur, India. His research interests include solar radiation estimation and forecasting, renewable energy, energy management, optimization techniques, and machine learning in power system. He has published 12 scientific papers in reputed journals, and international conference.